STA 431s13 Assignment Ten[1]

For the SAS question, please bring your log and list files to the quiz. Do not write anything on the printouts except your name and student number. The other questions are just practice for the quiz on Friday April 5th, and are not to be handed in.

1. Here is a factor analysis model in which all the observed variables are *standardized.* That is, they are divided by their standard deviations as well as having the means subtracted off. This gives them mean zero and variance one. Therefore, we work with a correlation matrix rather than a covariance matrix; that's the classical way to do factor analysis. Let

$$\begin{aligned} Z_1 &= \lambda_1 F_1 + e_1 \\ Z_2 &= \lambda_2 F_2 + e_2 \\ Z_3 &= \lambda_3 F_3 + e_3, \end{aligned}$$

where F_1, F_2 and F_3 are independent $N(0,1)$, e_1, e_2 and e_3 are normal and independent of each other and of F_1, F_2 and F_3, $V(Z_1) = V(Z_2) = V(Z_3) = 1$, and λ_1, λ_2 and λ_3 are nonzero constants. The expected values of all random variables equal zero.

   (a) What is $V(e_1)$? $V(e_2)$? $V(e_3)$?

   (b) What is $Corr(F_1, Z_1)$?

   (c) Give the communality of Z_j. Recall that the communality is the proportion of variance explained by the common factor(s). That is, it is the proportion of $Var(Z_j)$ that does not come from e_j.

   (d) Give the variance-covariance matrix (correlation matrix) of the observed variables.

   (e) Are the model parameters identifiable? Answer Yes or No and prove your answer.

   (f) Even though the parameters are not identifiable, the model itself is testable. That is, it implies a set of equality restrictions on the correlation matrix $\boldsymbol{\Sigma}$ that could be tested, and rejecting the null hypothesis would call the model into question. State the null hypothesis. Again, it is a statement about the $\sigma_{i,j}$ values.

2. Here is another factor analysis model. This one has a single underlying factor. Again, all the observed variables are standardized.

$$\begin{aligned} Z_1 &= \lambda_1 F + e_1 \\ Z_2 &= \lambda_2 F + e_2 \\ Z_3 &= \lambda_3 F + e_3, \end{aligned}$$

where $F \sim N(0,1)$, e_1, e_2 and e_3 are normal and independent of F and each other with expected value zero, $V(Z_1) = V(Z_2) = V(Z_3) = 1$, and λ_1, λ_2 and λ_3 are nonzero constants with $\lambda_1 > 0$.

[1] Copyright information is at the end of the last page.

(a) What is $V(e_1)$? $V(e_2)$? $V(e_3)$?

(b) Give the communality of Z_j.

(c) Write the reliability of Z_j as a measure of F. Recall that the reliability is defined as the squared correlation of the true score with the observed score.

(d) Give the variance-covariance (correlation) matrix of the observed variables.

(e) Are the model parameters identifiable? Answer Yes or No and prove your answer.

3. Suppose we added another variable to the model of Question 2. That is, we add

$$Z_4 = \lambda_4 F + e_4,$$

with assumptions similar to the ones of Question 2. Now suppose that $\lambda_2 = 0$.

(a) Is λ_2 identifiable? Justify your answer.

(b) Are the other factor loadings identifiable? Justify your answer.

(c) State the general pattern that is emerging here.

4. Suppose we added a fifth variable to the model of Question 3. That is, we add

$$Z_5 = \lambda_5 F + e_5,$$

with assumptions similar to the ones of Question 2. Now suppose that $\lambda_3 = \lambda_4 = 0$.

(a) Are λ_3 and λ_4 identifiable? Justify your answer.

(b) Are the other three factor loadings identifiable? Justify your answer.

(c) State the general pattern that is emerging here.

5. We now extend the model of Question 2 by adding a second factor. Let

$$\begin{array}{rcl}
Z_1 & = & \lambda_1 F_1 + e_1 \\
Z_2 & = & \lambda_2 F_1 + e_2 \\
Z_3 & = & \lambda_3 F_1 + e_3 \\
Z_4 & = & \lambda_4 F_2 + e_4 \\
Z_5 & = & \lambda_5 F_2 + e_5 \\
Z_6 & = & \lambda_6 F_2 + e_6,
\end{array}$$

where all expected values are zero, $V(e_i) = \omega_i$ for $i = 1, \ldots, 6$, $V(F_1) = V(F_2) = 1$, $Cov(F_1, F_2) = \phi_{12}$, the factors are independent of the error terms, and all the error terms are independent of each other. All the factor loadings are non-zero with $\lambda_1 > 0$ and $\lambda_4 > 0$.

(a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

6. In Question 5, suppose we added just two variables along with the second factor. That is, we omit the equation for Z_6. Are the model parameters identifiable in this case? Answer Yes or No; show your work.

7. Let's add a third factor to the model of Question 5. That is, we add

$$\begin{aligned} Z_7 &= \lambda_7 F_3 + e_7 \\ Z_8 &= \lambda_8 F_3 + e_8 \\ Z_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

with $\lambda_7 > 0$ and other assumptions similar to the ones we have been using. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

8. In this factor analysis model, the observed variables are *not* standardized, and the factor loading for D_1 is set equal to one. Let

$$\begin{aligned} D_1 &= F + e_1 \\ D_2 &= \lambda_2 F + e_2 \\ D_3 &= \lambda_3 F + e_3, \end{aligned}$$

where $F \sim N(0, \phi)$, e_1, e_2 and e_3 are normal and independent of F and each other with expected value zero, $V(e_1) = \omega_1, V(e_2) = \omega_2, V(e_3) = \omega_3$, and λ_2 and λ_3 are nonzero constants.

(a) Calculate the variance-covariance matrix of the observed variables.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

9. We now extend the preceding model by adding another factor. Let

$$\begin{aligned} D_1 &= F_1 + e_1 \\ D_2 &= \lambda_2 F_1 + e_2 \\ D_3 &= \lambda_3 F_1 + e_3 \\ D_4 &= F_2 + e_4 \\ D_5 &= \lambda_5 F_2 + e_5 \\ D_6 &= \lambda_6 F_2 + e_6, \end{aligned}$$

where all expected values are zero, $V(e_i) = \omega_i$ for $i = 1, \ldots, 6$,

$$V \left[\begin{array}{c} F_1 \\ F_2 \end{array} \right] = \left[\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array} \right],$$

and $\lambda_2, \lambda_3, \lambda_5$ and λ_6 are nonzero constants.

(a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done in Question 8.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

10. Let's add a third factor to the model of Question 9. That is, we add

$$\begin{aligned} D_7 &= F_3 + e_7 \\ D_8 &= \lambda_8 F_3 + e_8 \\ D_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

and

$$V \left[\begin{array}{c} F_1 \\ F_2 \\ F_3 \end{array} \right] = \left[\begin{array}{ccc} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{array} \right],$$

with $\lambda_8 \neq 0$, $\lambda_9 \neq 0$ and so on. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

11. The SAS part of this assignment is based on the Poverty Data. The data are given in the file poverty.data. There is a link on the course web page in case the one in this document does not work. This data set contains information from a sample of 97 countries. In order, the variables include Live birth rate per 1,000 of population, Death rate per 1,000 of population, Infant deaths per 1,000 of population under 1 year old, Life expectancy at birth for males, Life expectancy at birth for females, and Gross National Product per capita in U.S. dollars. There is also a categorical variable representing location (continent), and finally the name of the country. We won't use the last two columns of data.

This can be a very challenging and frustrating data set to work with, because correlated measurement errors produce negative variance estimates and other numerical problems almost everywhere you turn. To make your job easier, please confine your analyses to the following four variables:

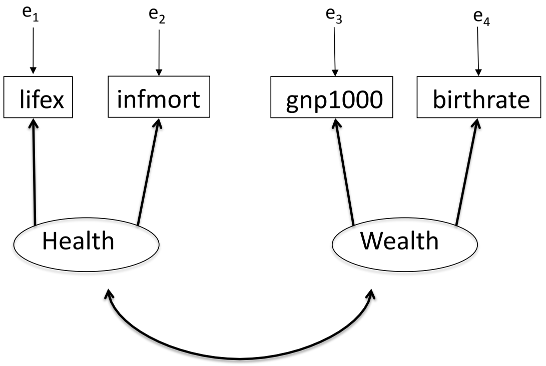
Life Expectancy: Average of life expectancy for males and life expectancy for females.

Infant mortality rate.

Birth rate.

GNP/1000 = Gross national product in thousands of dollars.

Here is a picture of a factor analysis model with 2 factors.



The reason for making birth rate an indicator of wealth is that birth control costs money.

(a) Fit the model with `proc calis`. Include the `pcorr` option, so you will get $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$. You will have to re-parameterize. Which of the two standard re-parameterizations should you choose? Suppose we are interested in the correlation between Health and Wealth.

(b) What is the parameter $\boldsymbol{\theta}$ for this model? *Give your answer in the form of a list of names from your SAS job.*

(c) Prove that the re-parameterization you have chosen results in a parameter vector that is identifiable – at least, it's identifiable in most of the parameter space.

(d) Does this model fit the data adequately? Answer Yes or No, and back up your answer with two numbers from the printout: The value of a test statistic, and a p-value.

(e) What is the maximum likelihood estimate of the correlation between factors? The answer is a single number from the printout.

(f) Now fit a model with the other common re-parameterization, again including the `pcorr` option.

   i. Compare the two likelihood ratio tests for model fit. What do you see?

   ii. Compare the two $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ matrices. What do you see?

   iii. Give the maximum likelihood estimate of λ_2/λ_1 based on output from the first model. Can you find this number in the output from the second model?

   iv. based on the output from the second model, give the maximum likelihood of the correlation between Health and Wealth. Can you find this number in the output from the first model?

(g) Finally, the high estimated correlation between factors from the first part of this question suggests that there might be just one underlying factor: wealth. Try a single-factor model and see if it fits. Locate the relevant chi-squared statistic, degrees of freedom and p-value. Do the estimated factor loadings make sense?

12. In a reaction time study, subjects are seated at a screen. A light flashes on the screen, and they press a key as fast as they can; the time between the light flash and the key press is recorded automatically.

    After some warmup trials, the subjects do the task 50 times, so 50 reaction times are recorded. The 50 times are divided randomly into two sets of 25, and then the median is calculated for each set. In the end, each subject produces two median reaction times.

    The scientists locate sample of university student volunteers whose parents and grandparents are also available to do the experiment. When all the data have been collected, there is a data file with n lines of data. Each line of data has 14 numbers. There are two median reaction times for each of the following individuals:

    - The student
    - Mother
    - Father
    - Maternal grandmother (mother's mother)
    - Maternal grandfather (mother's father)
    - Paternal grandmother (father's mother)
    - Paternal grandfather (father's father)

    As in most applications of statistical methods to real data, your job is to translate this flood of words into a statistical model.

    (a) Make a path diagram. Do not write any coefficients on the arrows yet. This means that your model could represent either an original model or a re-parameterized model. I believe that when a study of this type is done carefully, it is safe to assume that the errors are all independent. Are you making any other assumptions? (I did.)

    (b) Now, thinking about heredity the way you understand it, write coefficients on the straight arrows. Make your answer apply to the *original* model. Do not re-parameterize it yet. How many different coefficients do you need? Remember, a simpler model is better, as long as it's reasonable.

    (c) Write your original model in scalar form. You may assume that all the distributions are normal – pretty reasonable in this case. Include means for the latent variables, even though they are un-knowable. After all, grandparents are probably slowest on average.

    (d) Calculate the covariance matrix for your original model.

    (e) Now write a re-parameterized (surrogate) model with zero expected values, and other restriction(s) of your choice.

        i. Are the parameters of your new model identifiable? Answer Yes or No. If the answer is Yes, prove it. If the answer is No, fix u the model and try again.

ii. Does your new model impose exactly the same restrictions on the covariance matrix that the original model does? If so, in that sense it's *equivalent* to the original model, and it's a good re-parameterization.

(f) Remembering that the primary purpose of this study is to assess the role of heredity in reaction time, would the inclusion of everyone's age make the study better? Why or why not?

13. In a study of maternal behaviour in cats, mother cats with new litters of kittens were injected daily with estrogen, a female sex hormone. The cats were randomly assigned to different dosages (amounts) of estrogen. There are lots of different dosages, so dosage may be treated as a continuous variable. Because the exact amount injected is known, the variable Dosage is observed without error.

After three days, the amount of estrogen in the animal's bloodstream is measured, once. Of course it is measured with error. Then for the next seven days, the following maternal behaviours are recorded.

- Nursing time in total minutes.
- Licking in total number of times the cat licked one of her kittens.
- Retrievals: The mother cat picks up one of her kittens by the skin on the back of its neck, and carries it somewhere.

Again, your job is to come up with a reasonable model.

(a) Make a path diagram. Do not write any coefficients on the arrows yet. This means that your model could represent either an original model or a re-parameterized model.

(b) Write the equations of your original model in scalar form. Don't worry about expected values or distributions this time

(c) Calculate the covariance matrix for your original model.

(d) Now write a re-parameterized (surrogate) model with zero expected values, and other restriction(s) of your choice.

i. Are the parameters of your new model identifiable? Answer Yes or No. If the answer is Yes, prove it. If the answer is No, fix u the model and try again.

ii. Does your new model impose exactly the same restrictions on the covariance matrix that the original model does? If so, it's a good re-parameterization.
